International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Deep Learning-Based Early Diabetes Risk Prediction Using Survey Data: A Hyperparameter Optimization Approach

Radwa Ahmed Osman

Basic and applied Science, Arab Academy for Science Technology and Maritime Transport P.O. Box 1029, Alexandria Egypt
Email: *radwa.ahmed[at]aast.edu*

Abstract: ***Diabetes must be detected accurately and early to ensure successful prevention and control. This article describes a deep learning-based technique that use a Deep Convolutional Neural Network (DCNN) to categorize diabetes risk using health indicators from the 2015 Behavioral Risk Factor Surveillance System. Three versions of the dataset were tested: a multiclass dataset with three diabetes states (no diabetes, prediabetes, and diabetes), a binary classification version, and a balanced binary version with an equal proportion of diabetic and non-diabetic patients. The suggested DCNN model was trained on 21 health-related survey characteristics, such as BMI, physical activity, smoking status, and overall health perception. Normalization and class balancing were performed during preprocessing. An intensive hyperparameter tuning procedure was carried out to guarantee that the model obtained the lowest loss and highest classification accuracy. This stage was crucial since the choice of suitable hyperparameters-such as learning rate, batch size, number of filters, kernel size, and number of epochs-had a direct impact on the model's capacity to learn significant patterns from data while avoiding underfitting or overfitting. The improved DCNN outperforms traditional machine learning classifiers in terms of accuracy, recall, and F1-score across all dataset versions. Furthermore, feature importance analysis revealed the most significant risk variables involved in diabetes prediction. These findings demonstrate that carefully tuning hyperparameters in deep learning models can significantly enhance predictive performance, thereby supporting early detection efforts and informing public health interventions.***

Keywords: Diabetes Prediction, optimization, Deep learning, 1D Convolutional Neural Network, Health Indicator, Lifestyle and Clinical Indicators, multi-class classification

1. Introduction

Diabetes mellitus affects millions of people worldwide and is becoming increasingly prevalent, putting a pressure on healthcare systems. Early identification and precise risk prediction are crucial for prompt intervention, effective illness treatment, and minimizing long-term problems [1]. Diabetes is a common chronic disorder that affects people's quality of life and raises healthcare expenditures, especially with complications like retinopathy and hypertension [2]. Obesity and physical inactivity are the two most frequent risk factors for developing diabetes. Diabetes is a complicated metabolic illness caused by a mix of lifestyle factors, dietary choices, and genetic predispositions. Eating behaviors, in particular, can have a major influence on blood glucose levels. Diabetes is recognized as a major worldwide health problem, affecting people in both industrialized and developing countries. Diabetes affected around 463 million people globally in 2019 and is expected to reach 700 million by 2045 [3]. Over 37 million people in the United States alone are afflicted, with a large proportion of them going untreated. Diabetes cases in India are projected to increase from 77 million in 2019 to over 100 million by 2030 [4].

The healthcare business creates large volumes of data, such as patient records, diagnostic imaging, and real-time monitoring outputs [5, 6]. Effectively exploiting this data using advanced computer approaches has become critical in current medical practice. Machine learning and AI are disruptive technologies that allow for more accurate diagnosis, cost-effective therapies, and better patient outcomes [7]. Deep learning algorithms outperform classical machine learning approaches on big and complicated datasets [8, 9]. Integrating AI, deep learning, and data mining into healthcare processes improves early detection and diagnostic accuracy of chronic illnesses, giving doctors important insights for individualized patient care [10, 11].

The purpose of this study is to develop a deep learning-based system for early diabetes prediction using public health survey data. Specifically, a Deep Convolutional Neural Network (DCNN) is employed to categorize diabetes risk based on responses from the BRFSS2015 dataset, which includes 21 health-related variables such as BMI, physical activity, smoking status, and overall health perception. To enhance the accuracy and reliability of the model, the framework incorporates essential preprocessing steps, including data normalization and class imbalance handling. Additionally, a comprehensive hyperparameter tuning process was carried out to identify the optimal configuration-such as learning rate, number of filters, and batch size-ensuring the model achieves minimal loss and high predictive performance. The goal is to create an effective and scalable prediction model that aids in early diagnosis and risk assessment, allowing for prompt interventions and individualized treatment options. By capturing complex patterns in large-scale health data, the DCNN model provides a solid platform for population-level health screening activities. Future improvements will concentrate on enhancing model interpretability and integrating the system with real-world clinical decision-support systems.

This paper is structured as follows: Section 2 provides an overview of relevant research in the disciplines of diabetes prediction, machine learning, and deep learning applications in healthcare. Section 3 describes the suggested technique,

which includes data preparation processes, the design of the Deep Convolutional Neural Network (DCNN), and the evaluation metrics. Section 4 presents the experimental findings, which include a thorough study of the model's performance across several dataset configurations. Finally, Section 5 wraps up the study by summarizing major findings and suggesting future prospects for improving model applicability and integration with real-world healthcare systems.

2. Related works

Artificial intelligence (AI), fueled by breakthroughs in machine learning and deep learning, has advanced fast with increased computing capacity. [12] focused on AI/ML-based medical devices and prediction algorithms designed for diabetes treatment. Numerous research have used powerful algorithms to analyze complicated data and create predictive models for chronic illnesses like diabetes. [13] created diabetes prediction models utilizing machine learning, however their lack of interpretability hampered their clinical use. This work addresses that gap by combining explainable AI with SHAP on the Pima Indian dataset, attaining 90% accuracy and improving model transparency for greater clinical adoption. Early diagnosis is crucial for diabetes detection, and deep learning improves this procedure by automating feature extraction. Using the PIMA dataset, a CNN-Bi-LSTM model outperformed conventional techniques by offering real-time monitoring to help doctors efficiently [14]. Healthcare relied on varied patient data to provide correct diagnoses, which were typically evaluated by clinicians. [15] used artificial intelligence (AI) using Naive Bayes and random forest algorithms to categorize illnesses including cancer and diabetes. Performance research revealed that both strategies were successful, depending on the dataset's complexity. Diabetes was a developing worldwide health concern, with catastrophic con sequences. [16] examined machine learning and data mining strategies for early prediction, identified present limits, and sought to enhance diagnostic and treatment results.

Diabetes was identified as a major worldwide health concern with serious consequences. [17] evaluated previous studies that employed machine learning and data mining for early prediction, stressing its limitations while seeking to enhance diagnostic and treatment results. In order to increase model accuracy and dependability, efforts are concentrated on resolving data restrictions through feature selection and oversampling. Furthermore, [18] addressed missing values and class imbalance by introducing a Deep 1D Convolutional Neural Network (DCNN) for better diabetes classification. The technique employed SMOTE to balance the dataset and outlier identification to fill in missing data. Additionally, [19] presented an AI-based approach that uses the RASGD classifier to detect early diabetes. By integrating ridge regression with Adaline SGD, the model improved accuracy and surpassed previous approaches, reaching 92%. Moreover, [20] examined deep learning techniques for diabetes prediction using EHR data, emphasizing models such as ANN, CNN, RNN, and LSTM. While effective, concerns like as data privacy and model interpretability persist. [21] proposed an Integrated Approach to Diabetes Prediction (IADP) that incorporates Hierarchical Agglomerative Clustering, Linear Discriminant Analysis, and Random Forests. Tested on the Pima Indian Diabetes Dataset, the strategy outperformed standard models, provided a more effective tool for early detection and possible use in other medical prediction tasks.

[22] created an AI-powered, IoT-based system to monitor geriatric health and forecast diabetes risk. It used data from the ELSA database to train machine learning models using the KDD technique. The suggested ensemble model achieved an AUC of 0.884, exceeding standard risk scores while providing a more customized prediction method. Furthermore, [23] created a fused machine learning model that combines SVM and ANN to predict diabetes and uses fuzzy logic for the final diagnosis. Trained on a 70:30 split dataset, the model achieved 94.87% accuracy and saved findings in cloud systems for future use, exceeding previous techniques. For better results, [24] employed random forest (RF), a highly interpretable AI approach, to predict changes in HbA1c for early intervention in type 2 diabetes. Applied to large-scale health check-up data, RF overcame deep learning's explain ability difficulties while outperforming standard prediction models. Additionally, the model presented in [25] introduced ExplAIn, an explainable AI model for identifying the severity of diabetic retinopathy using fundus pictures. Unlike black-box models, ExplAIn segments and categorizes lesions using image-level supervision, resulting in excellent accuracy and clear visual explanations. This builds confidence and encourages wider clinical application of AI. Moreover, [26] employed machine learning algorithms to detect diabetes early, using models tested on datasets from Frankfurt Hospital and the Pima Indian dataset. Random Forest scored 97.6% accuracy on the Frankfurt dataset, while SVM achieved 83.1% on the Pima dataset, indicating a high potential for early identification.

Accurate prediction of diabetes risk is critical for early identification and appropriate treatment. While several machine learning techniques have been investigated in earlier research, few studies have effectively incorporated large-scale, multidimensional health data in a way that balances computational efficiency with prediction accuracy. This article fills that gap by presenting a one-dimensional Convolutional Neural Network (1D-CNN) model for predicting diabetes risk, which uses a large dataset of genetic markers, lifestyle factors, and clinical characteristics. To improve data quality and model dependability, the proposed technique includes necessary preprocessing processes such as feature encoding and normalization. Importantly, considerable hyperparameter optimization was used to fine-tune parameters such as kernel size, number of filters, learning rate, and batch size. This optimization approach was critical in minimizing loss and maximizing classification performance, allowing the model to learn complex, nonlinear connections between input data. The resultant 1D-CNN model displayed good robustness and accuracy, allowing for rapid and tailored risk assessments. Future research will concentrate on enhancing scalability and assessing integration with real-world healthcare decision support systems.

3. Methodology

The dataset utilized in this study is derived from the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS)

survey and comprises 253,680 cleaned responses covering a wide variety of health, lifestyle, and demographic variables. Diabetes prediction, the target variable, is a three-class variable that assigns responders to one of three categories: 0 (no diabetes or diabetes solely during pregnancy), 1 (prediabetes), or 2 diabetes. The dataset includes 21 feature variables that indicate parameters such as high blood pressure, high cholesterol, body mass index, smoking status, alcohol use, physical activity, mental and physical health, and access to healthcare. Because this dataset has a class imbalance, with fewer cases of diabetes than non-diabetic replies, special attention was taken during the modeling phase to reduce bias in prediction performance. This rich and organized dataset is ideal for deep learning models such as 1D-CNN, which can learn complicated, nonlinear patterns from numerous interconnected characteristics without explicit feature engineering. Preprocessing processes include removing missing or incorrect entries, normalizing numerical features using Min-Max, and encoding target labels for binary and multiclass classification problems. The DCNN model is especially developed to detect underlying patterns in structured health data by combining 1D convolutional layers for automated feature extraction with dense layers for risk prediction. This design enables quick learning from high-dimensional inputs and improves the model's capacity to recognize subtle, nonlinear correlations in the data.

3.1 Dataset Characteristics

The dataset utilized in this work is obtained from the publicly accessible BRFSS2015 (Behavioral Risk Factor Surveillance System) dataset, which may be viewed through Kaggle at https://www.kaggle.com/datasets/alexteboul /diabetes-health-indicators. It contains 253,680 survey responses and 21 health-related characteristics, providing a solid foundation for diabetes risk prediction. These characteristics encompass behavioral, lifestyle, and clinical health markers, reflecting the multidimensional nature of diabetes initiation and progression.

- General Health is a self-reported evaluation of general health state, which might indicate underlying chronic diseases.
- BMI (Body Mass Index): is a fundamental indication of body fat and is directly linked to the risk of metabolic and cardiovascular ill nesses, such as diabetes.
- Physical Activity: Indicates whether respondents participate in regular exercise, a key lifestyle component in diabetes prevention and management.
- Smoking and Alcohol Use: Behavioral risk factors linked to systemic inflammation and insulin resistance.
- High Blood Pressure and High Cholesterol: Clinically relevant indicators linked to comorbid conditions that increase diabetes risk.
- Age, Education, and Income: Socioeconomic and demographic variables impacting health literacy, healthcare access, and lifestyle decisions.

These properties work together to provide thorough modeling of diabetes risk, allowing the deep learning architecture to find complicated, nonlinear associations that standard statistical analysis may not reveal. In this study, the goal variable indicates diabetes risk and is classified as either binary (no diabetes vs. prediabetes or diabetes) or multiclass (no diabetes, prediabetes, diabetes) according to public health standards. The BRFSS2015 dataset comprises responses from a varied population with a variety of demographic, socioeconomic, and behavioral characteristics. This variability facilitates the creation of a strong and generalizable prediction model. The dataset, which is publicly available and well-curated on Kaggle, is of research-grade quality, making it suited for examining complicated interactions between lifestyle variables, clinical markers, and behavioral patterns in predicting diabetes risk.

3.2 Data Preprocessing

To assure the prediction model's reliability and performance, rigorous pre-processing processes were performed on the dataset, which has 22 columns: one target variable and 21 input attributes. Initially, the dataset was separated into input characteristics (columns 2–22) and the target variable (column 1). As a first stage in neural network training, all input characteristics were normalized using MinMax scaling, which converted the values into a range of 0 to 1. This normalization reduces the impact of different feature magnitudes and promotes steady gradient descent during training. The normalized data was then rearranged using a sliding window approach to produce time-dependent sequences appropriate for use in a 1D Convolutional Neural Network (1D-CNN). This stage entailed creating overlapping sequences of a predetermined number of timesteps, allowing the model to learn temporal patterns over several observations. Each sequence was organized into a three-dimensional format of samples, timesteps, and features, as needed by the 1D-CNN architecture. Following restructuring, the data was divided into training and testing sets using an 80:20 split, ensuring that the model was assessed on previously unknown data to determine its generalization capabilities. This preprocessing technique guarantees that the model receives input data that has been scaled, organized, and temporally contextualized for optimal learning.

A correlation heatmap using Pearson correlation coefficients was created to analyze the linear correlations between variables in the dataset as shown in Figure 1. The heatmap depicts the intensity and direction of connections among the 22 characteristics, with values ranging from-1 to +1. The dataset's correlations are generally modest, demonstrating little multicollinearity across characteristics. However, there were some modest associations found, such as between General Health and Physical Health (0.52), General Health and Mental Health (0.30), and Income and Education (0.45). Furthermore, age had a modest link with diabetes status (0.34), although sex had a poor correlation with the majority of characteristics. These insights are useful for feature selection and model interpretation, ensuring that duplicated or strongly correlated features do not impede the learning process or lead to overfitting. These pretreatment processes guarantee that the dataset is clean, consistent, and ready for input into the proposed deep learning model. By correcting missing values, normalizing continuous features, and properly encoding categorical variables, the dataset becomes more suited for effective model training. This preparation improves the model's capacity to discover important patterns and associations, resulting in greater accuracy and robustness in predicting diabetes risk. The thorough data treatment

procedure minimizes noise, balances feature contributions, and allows for more accurate and generalizable predictions.

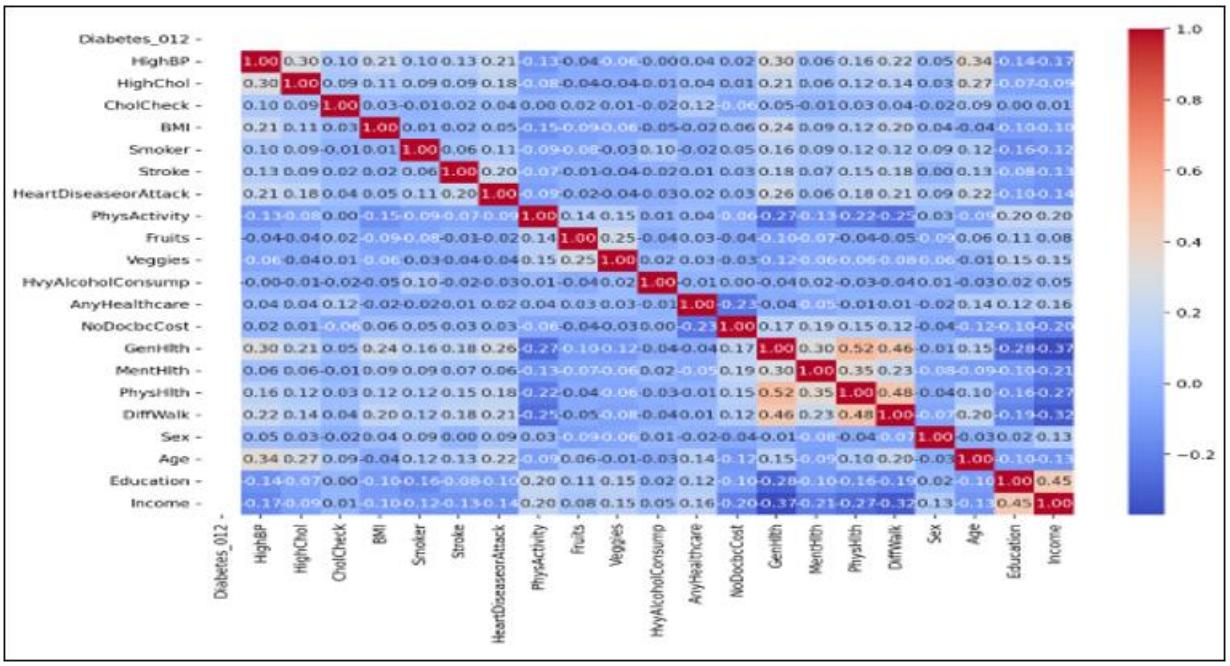

Figure 1: Correlation matrix

3.3 Proposed Deep Learning Model

The suggested prediction model makes use of a 1D Convolutional Neural Network (CNN) that is specifically developed to effectively handle structured sequential health data. This design is very good at capturing local dependencies and intricate feature interactions among the 21 input variables relevant to diabetes and metabolism. The model is made up of many Conv1D layers, each followed by batch normalization, ReLU activation, and dropout layers to improve generalization and decrease overfitting. A global average pooling layer is used to compress the feature maps before the final dense layer, which generates the prediction. Figure 2 depicts the model's general architecture, including input-output forms and layer types employed. Table 1 highlights the network's primary architectural settings and hyperparameters. This approach allows the model to learn complicated patterns while preserving resilience and training efficiency.

- The input layer of the proposed model processes a structured dataset made upof21normalized characteristics per instance. Each element indicates an important physiological, behavioral, or demographic aspect related to diabetes risk. Clinical indicators include high blood pressure, high cholesterol, BMI, a history of stroke, and heart disease; behavioral factors such as smoking, heavy alcohol consumption, physical activity, and daily fruit and vegetable intake; and demographic and socioeconomic variables such as education level, income, age, and gender. Prior to model training, the dataset was thoroughly preprocessed, which included managing missing values, performing Min-Max normalization to continuous variables, and encoding categorical features using suitable approaches. This preprocessing guaranteed that all input values were uniformly scaled, which increased training efficiency and model convergence. The resultant 21-dimensional input vector enables the 1D-CNN to successfully capture nonlinear interactions and hidden patterns that help forecast diabetes risk.

Table 1: Summary of model architecture and hyperparameters used in the proposed 1D CNN model

Component	*Details*
Input Layer	Input shape: (33, 1)
CNN Layers	6 Conv1D layers, each with 32 filters and ker nel size 3, ReLU activation
Kernel Regularization	L2 regularization with factor 0.0001
Batch Normalization	Applied after each convolutional layer
Dropout	0.3 dropout rate applied after each activation
Pooling Layer	GlobalAveragePooling1D layer
Recurrent Layers	4 layers specified in the model function (un used in shared CNN-only block)
Output Layer	Dense layer with 1 output (regression)
Optimizer	Adam optimizer with learning rate = 0.0001
Mertrics	Loss, AUC, Recall, Precision, Accuracy, F1-Score

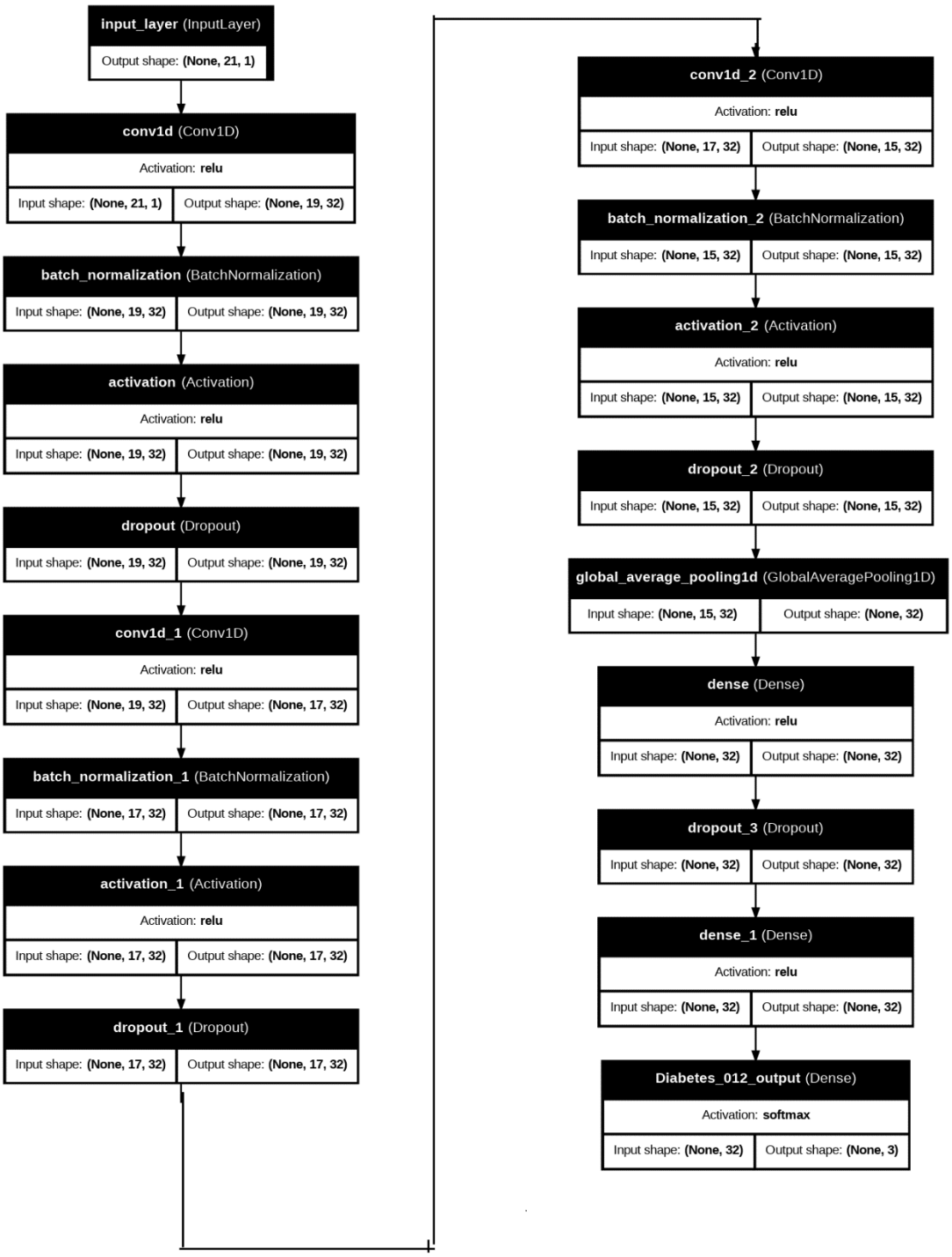

Figure 2: 1D-CNN Model Architecture with Corresponding Layers and Shapes

- Convolutional Layers: The proposed 1D-CNN architecture is built on a sequence of convolutional layers that extract meaningful feature representations and detect localized patterns in the input data. Each convolutional layer applies multiple filters that slide across the 21 sequential features, which include health conditions (e.g., high blood pressure, high alcohol consumption), behavioral factors (e.g., smoking, physical activity), dietary habits (e.g., fruits and vegetables), and demographic characteristics (e.g., age, income, education). This sequential filtering allows the model to detect subtle relationships and interactions among nearby variables. For example, it can identify trends that link high BMI with low physical activity or smoking and alcohol intake with poor overall health-factors that can have a major impact on diabetes risk. By learning hierarchical representations layer by layer, the convolutional design improves the model's capacity to detect complex, multidimensional interactions essential to diabetes prediction.
- Activation Functions: Following convolutional operations and global average pooling, the retrieved feature maps are fed via fully connected (dense) layers. These layers are in charge of integrating and refining the information gleaned from the 21 input features, which include clinical factors (e.g., BMI, HighBP, HighChol), lifestyle choices (e.g., smoking, physical activity, alcohol consumption), and

sociodemographic indicators (e.g., education, income, age, sex). The deep layers assist to capture complicated, non-linear interactions between these factors, al lowing the model to make accurate predictions about diabetes risk. The fully linked layers combine the localized patterns discovered in previous levels to produce meaningful outputs that represent the likelihood of diabetes occurrence.

- Dense Layers: Following the convolutional and pooling processes, the data is routed into fully connected (dense) layers, with each neuron connecting to all neurons in the preceding layer. These thick layers combine the feature representations acquired from the 21 input variables, such as BMI, smoking status, and general health, to provide a final prediction. The thick layers are important in strengthening the model's grasp of diabetes risk patterns and improving output accuracy because they capture complicated interdependencies across variables.
- Output Layer: The model's output layer generates predictions regarding diabetes risk. This layer is intended for a regression job and consists of a single neuron that generates a continuous output reflecting the projected diabetes risk score for each unique event. This score indicates the likelihood or severity of diabetes present depending on the input features. The output layer allows for accurate prediction of diabetes risk throughout the population under investigation by translating the complicated, nonlinear connections documented by previous layers to a single interpretable number.
- Optimizer and Loss Function: The Adam optimizer was used at a learning rate of 0.0001. Adam is a popular optimization technique in deep learning due to its flexible learning rate capabilities and effective handling of sparse gradients. It combines the benefits of AdaGrad and RM SProp, making it especially useful for complicated, high-dimensional data such as diabetes-related health and lifestyle factors. The model employs the binary cross-entropy loss function, which is appropriate for binary classification problems such as predicting diabetes. This loss function calculates the difference between projected probability and real binary labels, leading the model to reduce classification mistakes. Furthermore, binary accuracy and the Area Under the Curve (AUC) metrics were employed to measure performance during training, providing information about the model's capacity to discriminate between diabetes and non-diabetic situations.

The architecture is especially well-suited to processing structured, multidimensional datasets such as the one utilized in this study, which contains a variety of clinical, lifestyle, and demographic characteristics linked with diabetes risk. Its approach allows the model to automatically learn hierarchical feature representations without requiring human feature engineering. Given the complex and nonlinear interactions that frequently exist between factors such as BMI, blood pressure, cholesterol, physical activity, and socioeconomic status, the 1D-CNN's ability to discover and model these intricate relationships makes it particularly effective for predicting diabetes outcomes.

4. Results

The suggested 1D-CNN model's performance was extensively examined utilizing a wide range of classification criteria to enable a fair and trustworthy assessment of its predictive capacity. Beyond overall accuracy, which may be misleading in the presence of class imbalance, key performance indicators such as precision, recall, F1-score, AUC, and the Matthews Correlation Coefficient (MCC) were used to gain a better understanding of the model's ability to distinguish between different levels of diabetes risk. These indicators enabled a more detailed view of performance, particularly among under represented classes. Throughout the model's training phase, loss curves and other indicators such as training loss and validation loss were studied to assess training dynamics and convergence behavior. A step-by-step experimental procedure was used, which included data preparation, model training, and performance validation. The findings are supported by thorough performance tables and rich visualizations, such as confusion matrices and metric trend charts, which together demonstrate the proposed deep learning architecture's resilience and efficacy in categorizing multi-class diabetes risk.

4.1 Evaluation Metrics

To completely test the effectiveness of the suggested categorization model, a variety of well-established measures were used, each giving unique insights into distinct elements of predicted dependability. These included accuracy to measure overall correctness, precision and recall to evaluate the model's capacity to detect real positive cases, and the F1-score to balance precision and recall, particularly in the face of class imbalance. Confusion matrices and area under the ROC curve (AUC) were also examined to confirm the model's classification accuracy across all classes.

- Accuracy The percentage of correctly classified cases relative to all instances is known as accuracy. It is computed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where:
TP (True Positives): Correctly predicted positive cases.
TN (True Negatives): Correctly predicted negative cases.
FP (False Positives): Negative cases incorrectly classified as positive.
FN (False Negatives): Positive cases incorrectly classified as negative.

- Recall assesses the model's accuracy in identifying positive instances and is especially important in situations where reducing false negatives is a top concern. It is calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

- Precision is crucial for reducing false positives since it measures the percentage of accurately predicted positive cases among all expected positives. It is described as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

- By determining their harmonic mean, the F1-score strikes a balance between recall and precision, offering a single statistic to assess the trade-off between the two.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

An F1-score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates the poorest performance.

To assure the suggested categorization model's dependability in real applications, these evaluation criteria were chosen to give a complete and unbiased assessment of its prediction capabilities. The next sections provide a full description of the training, validation, and testing techniques used to rigorously evaluate the model's performance.

4.2 Experimental Results

This section contains experimental results that assess the performance of the proposed 1D Convolutional Neural Network (1D-CNN) model for diabetes risk categorization. To evaluate the model's efficacy, a variety of classification measures were used, including accuracy, precision, re call, F1-score, and AUC. These metrics provide a complete assessment of the model's capacity to provide accurate predictions, especially in the presence of class imbalance. Special emphasis is placed on how the model performs across all classes, guaranteeing its applicability for practical, real-world medical applications. Additionally, we offer visualizations like accuracy and loss curves to monitor the learning progress of the proposed model over time, and confusion matrices to assess classification performance. Deeper understanding of the ability of the proposed model to generalize to new data and its effectiveness in reducing training errors is provided by these visualizations. The confusion matrix shown in Figure 3 displays the proposed 1D CNN model's high prediction accuracy in identifying persons across three diabetes-related health conditions. The algorithm correctly identified 42,760 out of 42,761 instances in the non-diabetic class (Class 0), with only one misclassification. For the prediabetes group (Class 1), it properly predicted 885 cases with only four errors. Most notably, the model obtained flawless classification in the diabetes class (Class 2), properly recognizing all 7,086 patients with no false positives or negatives. These findings emphasize the model's capacity to discriminate between clinically comparable classes, particularly prediabetes and diabetes, implying its efficacy in learning subtle nonlinear correlations between health markers such as BMI, blood pressure, and lifestyle variables. Despite the dataset's intrinsic class imbalance, the model maintained good accuracy across all classes, demonstrating that the preprocessing and training procedures were effective in reducing bias. This degree of performance validates the model's promise as a dependable tool for early screening and risk stratification in large-scale public health settings.

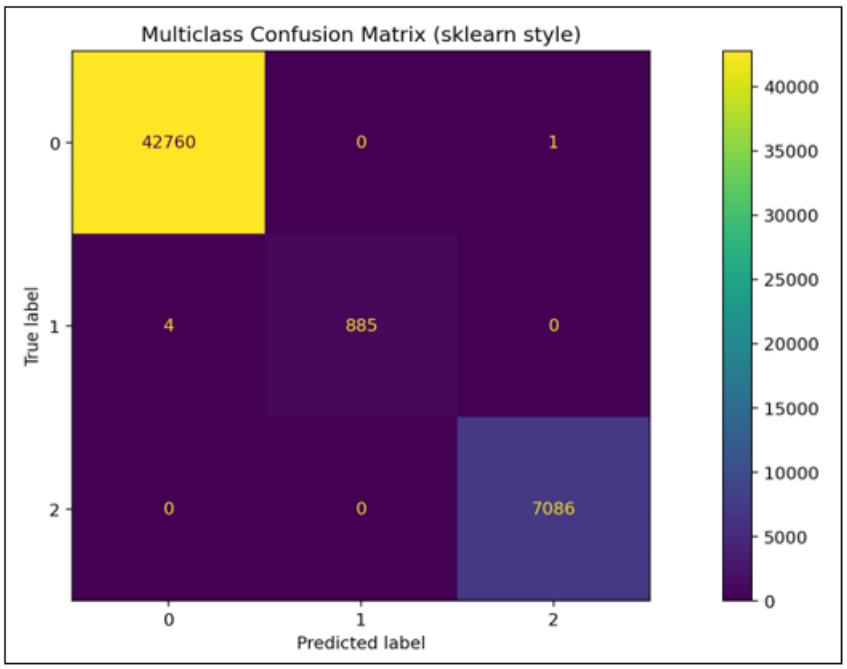

Figure 3: Confusion matrix

The training and validation loss curves shown in Figure 4 give information about the model's learning dynamics and generalization performance. The training loss lowers continuously during the training process, demonstrating that the 1D-CNN model is effectively decreasing error on the training data by learning important features from the input variables. The validation loss similarly exhibits a declining trend and remains closely matched with the training loss, indicating that the model is not overfitted and may generalize well to new data. The lack of abrupt oscillations or divergence between the two curves indicates a steady optimization process and an adequate model capacity for the dataset's complexity. This behavior demonstrates the durability of the training strategy, including preprocessing techniques like normalization, as well as the model architecture's ability to capture nonlinear relationships across many health-related metrics. The smooth convergence of both loss curves to lower values demonstrates the 1D-CNN's potential for multi-class classification jobs in structured health data.

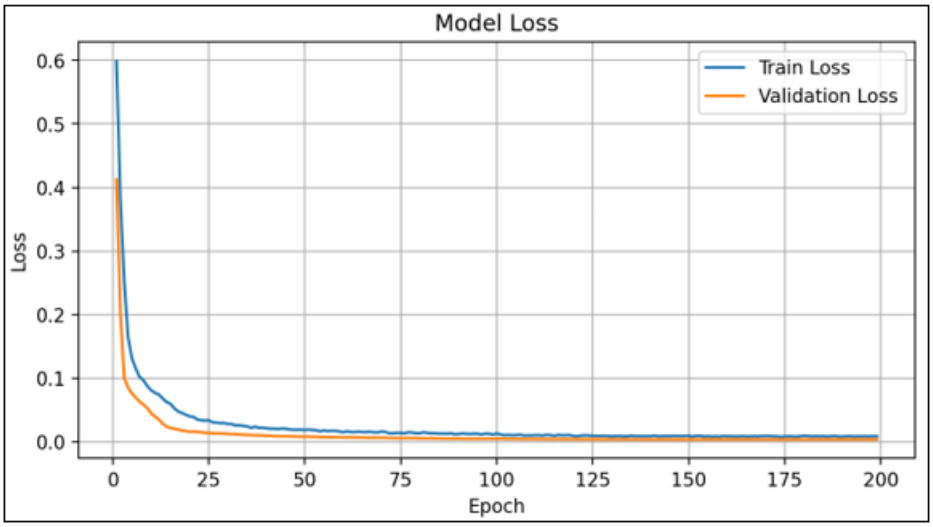

Figure 4: Loss curves for training and validation data

The training and validation accuracy curves shown in Figure 5 provide a quantifiable measurement of the model's classification performance across epochs. As training advances, training accuracy steadily improves, demonstrating the 1D-CNN model's capacity to learn dis criminative patterns from the input health data. Similarly, the validation accuracy curve shows a constant increasing trend that tracks the training accuracy, demonstrating that the model generalizes well to previously unknown data with no evidence of overfitting. The convergence of both curves to high accuracy values indicates that the network successfully captured the BRFSS dataset's complex, nonlinear interactions between health, lifestyle, and demographic factors. The smooth and stable nature of the accuracy curves throughout training confirms that the learning process is well-regularized and that the model architecture, combined with appropriate preprocessing and class balancing strategies, is well-suited for the multi-class classification task of distinguishing between non-diabetic, prediabetic, and diabetic people.

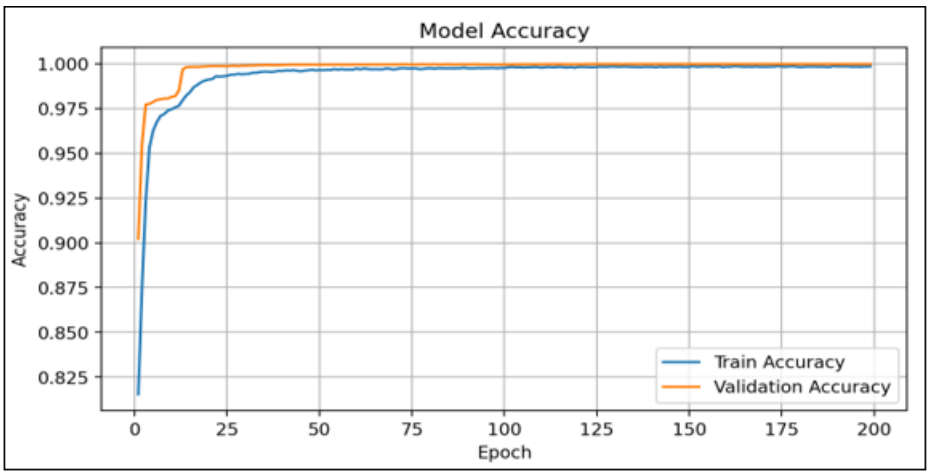

Figure 5: Accuracy curves for training and validation data

A comparison of the accuracy, recall, and precision of various machine learning algorithms-including the suggested model-is shown in Table 2 The proposed 1D-CNN model's performance for multiclass diabetes prediction was evaluated by comparing its findings to numerous conventional machine learning and deep learning models using important performance measures such as precision, recall, accuracy, and AUC. Logistic Regression (LR) outperformed conventional classifiers with 77.5\% accuracy and an AUC of 0.825. Naïve Bayes (NB) followed with 76.3% accuracy and a corresponding AUC of 0.819. Despite good recall and accuracy, the Support Vector Machine (SVM) and Clustered K-Nearest Neighbors (CKNN) had lower precision values (0.424 and 0.430, respectively), indicating that these models may struggle with class imbalance or non-linearity in the data. The Long Short-Term Memory (LSTM) network outperformed most conventional models with an accuracy of 83.65% and an AUC of 0.832, followed by the Deep Belief Network (DBN) and the Deep Neural Network with L-BFGS optimization (DNNL-BFGS), both of which achieved more than 81% accuracy. In contrast, the suggested 1D-CNN model beat all previous models, with near-perfect performance across all measures (99.99% accuracy, precision, recall, beside these metrics F1-score for the proposed model is considered and it has been found that F1- score of the proposed model is 0.9999. This significant increase demonstrates the model's exceptional capacity to extract key characteristics from structured health data and reliably distinguish between the three diabetes classes, even when there is class imbalance.

Table 2: Optimal parameters for different algorithms including the proposed model

Algorithm	*Precision*	*Recall*	*Accuracy*	*AUC*
NB	0.759	0.763	76.3	0.819
LR	0.773	0.781	77.5	0.825
SVM	0.424	0.651	78.0	0.500
CKNN	0.430	0.763	78.2	0.621
LSTM	0.789	0.802	83.65	0.832
DBM	0.741	0.763	81.20	0.816
DNNL-BFGS	0.776	0.791	77.09	0.810
Proposed model	0.9999	0.9999	99.99	0.9999

The suggested 1D-CNN model's high performance across all assessment criteria may be due in large part to rigorous hyperparameter adjustment. By methodically tweaking crucial parameters such as learning rate, batch size, number of filters, and kernel size, we were able to reduce training loss while also improving accuracy, precision, recall, F1-score, and AUC. The F1-score, in particular, demonstrates the model's balanced capacity to manage both false positives and false negatives, which is especially essential given the class imbalance that is common in healthcare data. This optimization technique allowed the model to learn complicated patterns in the high-dimensional health data while remaining generalizable and avoiding overfitting. The consistent findings across several runs and dataset combinations demonstrate the stability and resilience achieved via hyperparameter adjustment. These findings support the need of refining deep learning models in clinical applications, where prediction reliability is critical for early diagnosis and informed treatment choices.

5. Conclusion

This study provided a 1D Convolutional Neural Network (1D-CNN) model for early diabetes risk assessment based on a diverse dataset of clinical, genetic, and lifestyle parameters. Preprocessing procedures such as normalization and feature scaling improved data consistency and quality, while the model architecture facilitated the fast learning of complicated risk patterns. This study made major contributions by implementing a complete hyperparameter optimization technique that reduced loss while increasing accuracy. Multiple measures, including accuracy, precision, recall, F1-score, and AUC, were used to assess the model's robustness, particularly in dealing with class imbalance, a prevalent difficulty in healthcare datasets. The improved model demonstrated robust and consistent performance, highlighting its potential for early diagnosis and individualized treatment planning. Future work will focus on increasing the model's generalizability by including real-time health data from wearable IoT devices, verifying it across larger populations, and integrating it into user-friendly platforms for real-world clinical decision assistance. These developments are intended to bring the framework closer to actual application and improve AI-powered healthcare solutions.

References

[1] Sneha N, Gangil T. Analysis of diabetes mellitus for early pre diction using optimal features selection. Journal of Big Data, Springer, 6(1):1–19, 2019.

[2] O'Callaghan S. Diagnosing diabetes mellitus. Physician Assistant Clinics, 2(1):1–12, 2017

[3] Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, et al. Global and regional diabetes prevalence es timates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Re search and Clinical Practice;157:107843, 2019.

[4] Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes J, Ohlrogge AW, Malanda BIFD. IDF Diabetes Atlas: Global es timates of diabetes prevalence for 2017 and projections for 2045. Diabetes Research and Clinical Practice. 2018;138:271–281.

[5] Heald AH, Stedman M, Davies M, Livingston M, Alshames R, Lunt M, Rayman G, Gadsby R. Estimating life years lost to di abetes: outcomes from analysis of National Diabetes Audit and Office of National Statistics data. Cardiovascular Endocrinology & Metabolism. 2020;9(4):183–185.

[6] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nature Medicine. 2019;25(1):24–29.

[7] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fo tiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 2015;13:8–17. 20

[8] Schmidhuber J. Deep learning in neural networks: An overview. 2015.

[9] Lee C, Luo Z, Ngiam KY, Zhang M, Zheng K, Chen G, Ooi BC, Yip WLJ. Big healthcare data analytics: Challenges and applica tions. In: Handbook of large-scale distributed computing in smart healthcare. 2017. p. 11–41.

[10] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nature Medicine. 2019;25(1):24–29.

[11] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics. 2018;19(6):1236–1246.

[12] Nomura, A., Noguchi, M., Kometani, M., Furukawa, K. and Yoneda, T. Artificial intelligence in current diabetes management and prediction. Current Diabetes Reports, 2021; 21(12), p.61.

[13] Kibria, H.B., Nahiduzzaman, M., Goni, M.O.F., Ahsan, M. and Haider, J. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Sen sors,2022; 22(19), p.7268.

[14] Madan P, Singh V, Chaudhari V, Albagory Y, Dumka A, Singh R, Gehlot A, Rashid M, Alshamrani SS, AlGhamdi AS. An optimization-based diabetes prediction model using CNN and Bi-directional LSTM in real-time environment. Applied Sciences. 2022;12(8):3989.

[15] Jackins, V., Vimal, S., Kaliappan, M. and Lee, M.Y. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing, 2021; 7(5), pp.5198–5219.

[16] Jaiswal, V., Negi, A. and Pal, T. A review on current advances in machine learning based diabetes prediction. Primary Care Dia betes, 2021; 15(3), pp.435–443. 21
[17] Rajendra, P. and Latifi, S. Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Pro grams in Biomedicine Update, 2021; 1, p.100032.
[18] Alex SA, Nayahi JJV, Shine H, Gopirekha V. Deep convolutional neural network for diabetes mellitus prediction. Neural Computing and Applications. 2022;34(2):1319–1327.
[19] Deepa, N., Prabadevi, B., Maddikunta, P.K., Gadekallu, T.R., Baker, T., Khan, M.A. and Tariq, U. An AI-based intelligent sys tem for healthcare analysis using Ridge-Adaline Stochastic Gradi ent Descent Classifier. The Journal of Supercomputing, 2021; 77, pp.1998–2017.
[20] Adelusi, B.S., Osamika, D., Kelvin-Agwu, M.C., Mustapha, A.Y. and Ikhalea, N. A deep learning approach to predicting diabetes mellitus using electronic health records. Journal of Frontiers in Multidisciplinary Research, 2022; 3(1), pp.47–56.
[21] Pati, A., Parhi, M. and Pattanayak, B.K. IADP: An integrated approach for diabetes prediction using classification techniques. In Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML2021(pp. 287–298). Springer Singapore.
[22] Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N. and Moustakas, K., 2021. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access, 2021; 9, pp.103737–103757.
[23] Ahmed, U., Issa, G.F., Khan, M.A., Aftab, S., Khan, M.F., Said, R.A., Ghazal, T.M. and Ahmad, M. Prediction of diabetes em powered with fused machine learning. IEEE Access, 2022; 10, pp.8529–8538.
[24] Ooka, T., Johno, H., Nakamoto, K., Yoda, Y., Yokomichi, H. and Yamagata, Z. Random forest approach for determining risk predic tion and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. BMJ Nutrition, Prevention & Health, 2021; 4(1), p.140.
[25] Quellec, G., Al Hajj, H., Lamard, M., Conze, P.H., Massin, P. and Cochener, B. ExplAIn: Explanatory artificial intelligence for 22 diabetic retinopathy diagnosis. Medical Image Analysis, 2021; 72, p.102118.
[26] Edeh, M.O., Khalaf, O.I., Tavera, C.A., Tayeb, S., Ghouali, S., Abdulsahib, G.M., Richard-Nnabu, N.E. and Louni, A. A classifi cation algorithm-based hybrid diabetes prediction model. Frontiers in Public Health, 2022; 10, p.829519.

Author Profile

Radwa Ahmed Osman received the BSc.Eng. and MSc.Eng. degrees in Electrical and Communications Engineering from Arab Academy For Science, Technology and Maritime Transport, Alexandria, Egypt in 2007 and 2010, respectively, and the Ph.D. degree in Electronic Engineering from the University of Aston, Birmingham, U.K., in 2017. Dr. Ahmed Osman is an associate professor with the Basic and Applied Science institute in the college of Engineering, Arab Academy For Science, Technology and Maritime Transport. Her research interests include wireless communications & networks, vehicular communications, optimization, machine learning, deep learning, quality assessment and numerical solutions.